The Need for Earth Science Data Analytics
What are Your Analytics Requirements?

Earth Science Data Analytics Cluster

Steve Kempler, Moderator

January 8, 2016

ESIP Federation Meeting
Washington, DC

Session Focus

Session Focus:

- The ESDA Cluster (for new participants)
- What we have accomplished
- What we have done recently (where we are)
- What we still need to do

Obligatory Background Information

Earth Science Data Analytics (ESDA) Cluster Goal:

To understand where, when, and how ESDA is used in science and applications research through speakers and use cases, and determine what Federation Partners can do to further advance technical solutions that address ESDA needs. Then do it.

Ultimate Goal:

To Glean Knowledge about Earth from All Available Data and Information

Motivation

Increasing Amounts of Heterogeneous Datasets being made available to advance science research

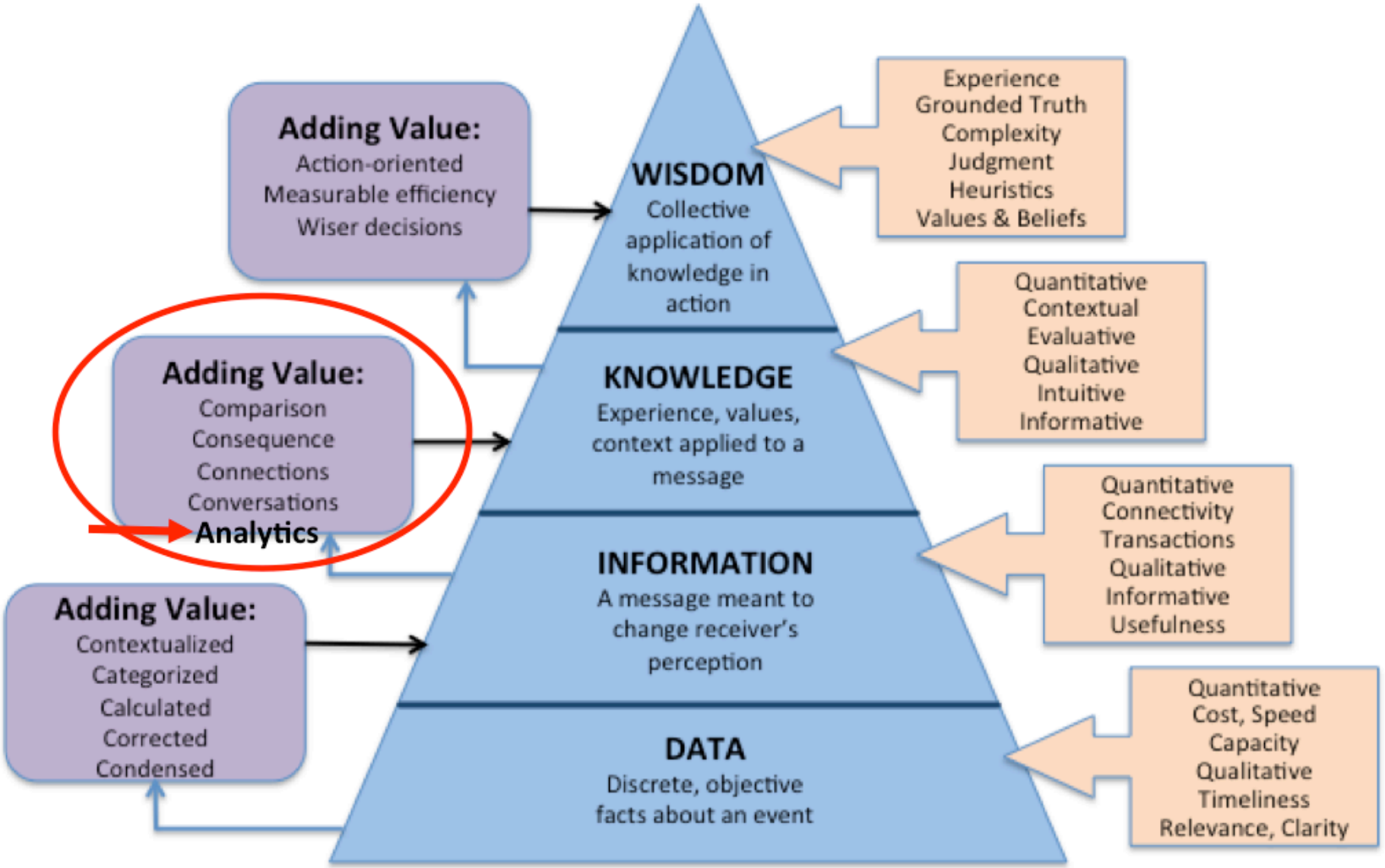
… and a lot of people/directives are addressing it

Thus, it is not necessarily about Big Data, itself.

It is about the **ability to examine large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information.**

That is:

To glean knowledge from data and information



(Adapted from: https://km4meu.wordpress.com/tag/dikw-pyramid/)

ESDA Cluster – What we have done

- 18 Telecons
- 7 face-to-face sessions
- 16 ‘guest’ presentations
- Created the ESDA specific use case template
- Gathered 18 use Cases
- Defined Earth Science Data Analytics (submitted for ESIP adoption)
- Specified 3 types of ESDA definition types
- Defined 10 Earth science data analytics goals (submitted for ESIP adoption)
- Commenced ESDA Tools/Techniques requirements analysis
  - Began gathering and describing known tools/techniques
  - Began analyzing use case ESDA tools/techniques usage/needs
- Presented our work at AGU

Earth Science Data Analytics Definition

The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

This encompasses:

- **Data Preparation** – Preparing heterogeneous data so that they can be jointly analyzed
- **Data Reduction** – Correcting, ordering and simplifying data in support of analytic objectives
- **Data Analysis** – Applying techniques/methods to derive results

Earth Science Data Analytics Goals

(read: Earth science data analytics needed ...)

1. To calibrate data
2. To validate data (note it does not have to be via data intercomparison)
3. To assess data quality
4. To perform coarse data preparation (e.g., subsetting data, mining data, transforming data, recovering data)
5. To intercompare datasets (i.e., any data intercomparison; Could be used to better define validation/quality)
6. To tease out information from data
7. To glean knowledge from data and information
8. To forecast/predict/model phenomena (i.e., Special kind of conclusion)
9. To derive conclusions (i.e., that do not easily fall into another type)
10. To derive new analytics tools

Data Analytics Goals

Why is it important to identify Data Analytics Goals

To better identify key needs that tools/techniques can be developed to address.

Basically, once we can categorize different goals of Data Analytics, we can better associate existing and future Data Analytics tools and techniques that will help solve particular problems.

Use Cases (gathered so far) Mapped to ESDA Goals

Use Cases	Earth Science Data Analaytics Goals									
	1	2	3	4	5	6	7	8	9	10
1 MERRA Analytics Services: Climate Analytics-as-a-Service										√
2 MUSTANG QA: Ability to detect seismic instrumentation problems			√	√				√		
3 Inter-calibrations among datasets	√	√			√					
4 Inter-comparisons between multiple model or data products					√					
5 Sampling Total Precipitable Water Vapor using AIRS and MERRA		√			√					
6 Using Earth Observations to Understand and Predict Infectious Diseases								√	√	
7 CREATE-IP - Collaborative REAnalysis Technical Environment - Intercomparison Project					√					
8 The GSSTF Project (MEaSUREs-2006)						√				
9 Science- and Event-based Advanced Data Service Framework at GES DISC					√					√
10 Risk analysis for environmental issues								√		
11 Aerosol Characterization					√				√	
12 Creating One Great Precipitation Data Set From Many Good Ones						√				
13 Reconstructing Sea Ice Extent from Early Nimbus Satellites	√			√						
14 DOE-BER AmeriFlux and FLUXNET Networks *						√			√	
15 DOE-BER Subsurface Biogeochemistry Scientific Focus Area *								√		
16 Climate Studies using the Community Earth System Model at DOE's NERSC center *								√	√	√
17 Radar Data Analysis for CReSIS *						√				
18 UAVSAR Data Processing, Data Product Delivery, and Data Service *						√				

* - Borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php]

Deriving Earth Science Data Analytics Requirements

Goal oriented Earth Science Data Analytics (ESDA) reveal requirements for needed data analytics tools/techniques

Motivation
How can we maximize the usability of large heterogeneous datasets to glean knowledge out of the data?

Earth Science Data Analytics: Definition
The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.

Methodology
Categorize/Analyze ESDA use cases; derive data analytics requirements; associate tools/techniques; perform gap analysis

Data Preparation | **Data Reduction** | **Data Analysis**

Earth Science Data Analytics: Goals

- To validate data
- To perform coarse data preparation
- To calibrate data
- To assess data quality
- To intercompare datasets
- To tease out information
- To glean knowledge
- To derive conclusions
- To forecast/ predict/model
- To derive new analytics tools

Earth Science Data Analytics: Initial Requirements

- Ingest from various sources; Homogenize data; Visualization; Sampling; Gridding
- Access large datasets; High speed processing; Subsetting, mining, machine learning
- Homogenize data; Intercomparison statistics; Pattern recognition
- Seek heterogeneous data relationships; Ingest from various sources; Image processing
- *Looking for Community input*
- Data exploration; Filter, mine, fuse, interpolate data; Manage custom code
- Ingest from various sources; High speed processing; Math functions
- Access large datasets; Assess erroneous data; Detect data anomalies
- Data exploration; Neural networks; Math/ Stat modeling; Near Real Time data
- Access very large datasets; homogenize data; visualization

Earth Science Data Analytics: Exemplary Tools, Techniques, Integrated Systems

Types of Analytics	Tools	Techniques		Integrated Systems
• Data Preparation • Data Reduction • Data Analysis	• R, SAS, Python, Java, C++ • SPSS, MATLAB, Minitab • CPLEX, GAMS, Gauss • Tableau, Spotfire • VBA, Excel, MySQL • Javascript, Perl, PHP • Open Source Databases • PIO, NCL, Parallel NetCDF • AWS, Cloud Solutions, Hadoop • MPI, GIS, ROI-PAC, GDAL	• Statistics functions • Machine Learning • Data Mining • Natural Language Processing • Linear/Non-linear Regression • Logical Regression • Time Series Models • Clustering • Decision Tree	• Factor Analysis • Principal Component Analysis • Neural Networks • Bayesian Techniques • Text Analytics • Graph Analytics • Visual Analytics • Map Reduce	• EarthServer (http://www.earthserver.eu) • NASA Earth Exchange (https://nex.nasa.gov/nex/) • EDEN (http://cda.ornl.gov/projects/eden/#) • EARTHDATA (https://earthdata.nasa.gov) • Giovanni (http://giovanni.gsfc.nasa.gov/giovanni/)

Compiled from: http://practicalanalytics.co/predictive-analytics-101/ and

Earth Science Data Analytics: Enabling Organizations

Research Data Sharing without barriers

Federation of Earth Science Information Partners
Fostering connections to make data matter

National Institute of Standards and Technology

OGC® Making location count.

The good news…

Earth Science Data Analytics: Preparing for the Future

Central England NERC Training Alliance — CENTA

Big data analysis to fuel environmental research at Reading University

2nd Annual Graduate Workshop on Environmental Data Analytics
July 27-31, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

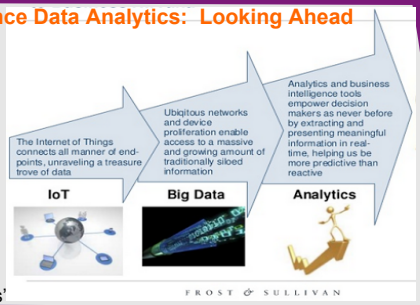
… offering degrees in Data Science

… summer school on Big Data Analytics

… online master's degree in data analytics

Earth Science Data Analytics: Looking Ahead

- Complete Gap Analysis between ESDA requirements and current tools/ technologies
- Continue to evolve tools/ techniques to address growing scope of the 'Internet of Things'

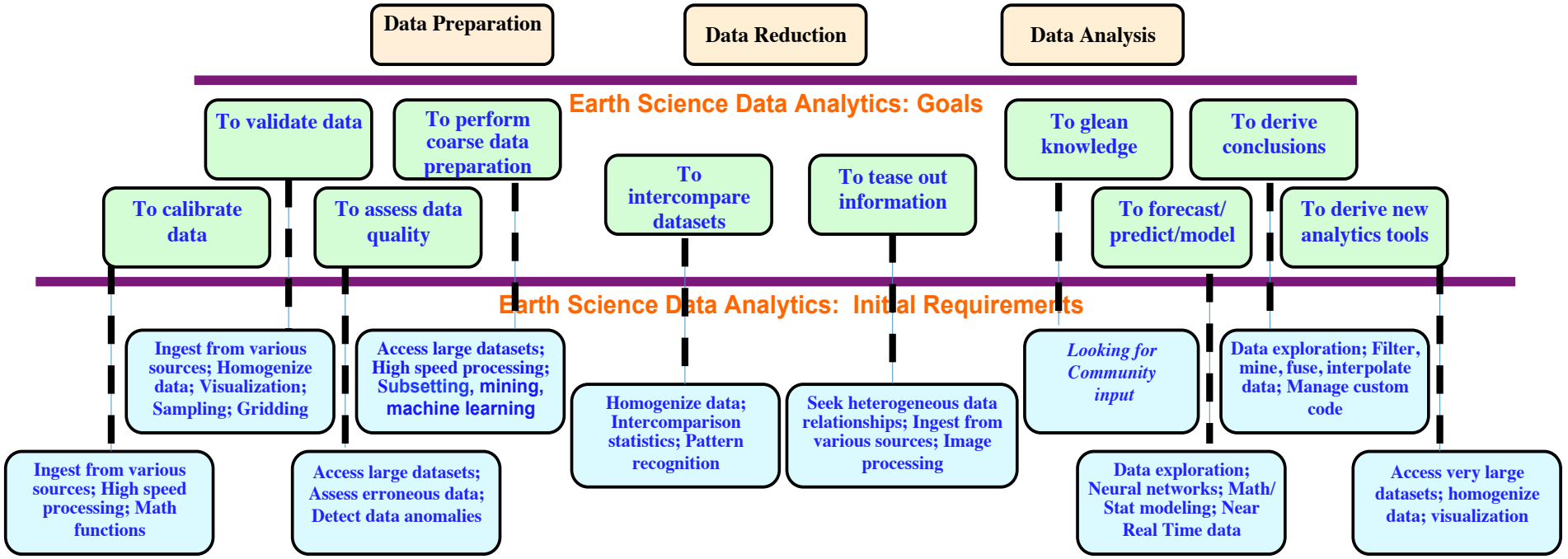


* Thanks to the work of the Earth Science Information Partners (ESIP) Federation, Earth Science Data Analytics (ESDA) cluster

Deriving Earth Science Data Analytics Requirements

Earth Science Data Analytics: Definition

The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data using a variety of data types to uncover patterns, correlations and other information, to better understand our Earth.



Earth Science Data Analytics
Exemplary Tools, Techniques, Integrated Systems

Types of Analytics	Tools	Techniques		Integrated Systems
• Data Preparation • Data Reduction • Data Analysis	• R, SAS, Python, Java, C++ • SPSS, MATLAB, Minitab • CPLEX, GAMS, Gauss • Tableau, Spotfire • VBA, Excel, MySQL • Javascript, Perl, PHP • Open Source Databases • PIO, NCL, Parallel NetCDF • AWS, Cloud Solutions, Hadoop • MPI, GIS, ROI-PAC, GDAL	• Statistics functions • Machine Learning • Data Mining • Natural Language Processing • Linear/Non-linear Regression • Logical Regression • Time Series Models • Clustering • Decision Tree	• Factor Analysis • Principal Component Analysis • Neural Networks • Bayesian Techniques • Text Analytics • Graph Analytics • Visual Analytics • Map Reduce	• EarthServer (http://www.earthserver.eu) • NASA Earth Exchange (https://nex.nasa.gov/nex/) • EDEN (http://cda.ornl.gov/projects/eden/#) • EARTHDATA (https://earthdata.nasa.gov) • Giovanni (http://giovanni.gsfc.nasa.gov/giovanni/)

So, where are we…

- √ Finalize ESDA Definition and Goal categories
- √ Write letter to ESIP Executive Committee proposing that the ESDA Definition and Goals be ESIP approved
- √ Characterize current use cases by Goal categories and other analytics driving considerations
- √ Derive requirements from use cases (still needs work) *
- Further validate requirements with (many) more additional use cases
- √ Survey/Describe existing data analytics tools/techniques *
- Perform gap analysis between ESDA requirements and available tools *
- Engage ESIP group interested in 'Emerging Big Data Technologies for Geoscience'
- Write our paper describing ... all the above

* Today's focus

We Began Describing Identified Tools/Techniques/Integrated Systems

Tool/Technique/ Integrated System	Description	Author
R	R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. (Wikipedia)	Steve
SAS	SAS (Statistical Analysis System) is a software suite developed by SAS Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics. SAS was developed at North Carolina State University from 1966 until 1976, when SAS Institute was incorporated. (Wikipedia)	Steve
Python	Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale. (Wikipedia)	Sean
Java		Steve
C++		Steve
SPSS		Sean
MATLAB		Sean
Mintab		Steve
CPLEX		Steve
GAMS		Steve
Gauss		Steve
Tableau	A tool that enables data visualization using a drag and drop interface.	Thomas
Spotfire	A tool that enables data mining and visualization of very large data sets. Similar to Excel but apparently easier to use for large data sets.	Thomas
VBA	(Visual Basic for Applications) An implementation of Visual Basic that enables user defined functions and interaction with Windows API and libraries.	Thomas
Excel	A spreadsheet program created by Microsoft that enables data analysis and visualization. It includes VBA.	Thomas

We Began Describing Identified Tools/Techniques/Integrated Systems

MySQL		Thomas
Javascript	A high level interpreted language used by most websites and browsers.	Thomas
Perl	A high level interpreted scripting language frequently used on UNIX computers. It is frequently used to wrap other programs together.	Thomas
PHP	A scripting language designed for web development. It can be used to create CGI (Common Gateway Interface) executable for web pages.	Thomas
Open Source Databases		Steve
PIO		Steve
NCL		Steve
Parallel NetCDF		Steve
AWS		Steve
Cloud Solutions		Steve
Statistics functions		-
Machine Learning	Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.	Chung-Lin
Data Mining	Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets (“big data”) involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.	Chung-Lin
Natural Language Processing	atural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.	Chung-Lin

We Began Describing Identified Tools/Techniques/Integrated Systems

Linear/Non-linear Regression	In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable Y (e.g., a sounding temperature) and one or more explanatory variables (or independent variables) denoted X, (or X1, X2...) (e.g., the satellite retrieved temperature(s)). The case of one explanatory variable is called simple linear regression. In statistics, nonlinear regression is a form of regression analysis in which observational data (e.g., Y) are modeled by a function which is a nonlinear combination of the model parameters (e.g., aX + bX2 +….) and depends on one or more independent variables (e.g., X or X1, X2,….). The data are fitted by a method of successive approximations.	Chung-Lin
Logical Regression		Bob
Time Series Models	Time Series Models are used to represent trends, often graphically, by applying temporal measurements within a sequence.	Bob
Clustering	Clustering is an approach to organize objects into a classification and can be accomplished utilizing various methods, including statical techniques.	Bob
Decision Tree	A Decision Tree is a graphical representation of the sequence of decisions to be completed when answering a particular question.	Bob

Then We Discovered…

"The Field Guide to DATA SCIENCE", Booz/Allen/Hamilton, 2015
(Thanks Ethan)

- This opened our eyes to a great resource that associates computational techniques to specific data science 'stages':
  - Describe, Discover, Predict, Advise
    - These stages are describe in terms of increasing maturity
    - Interpreted for Earth science, each stage would have independnet maturity levels. We would call them 'goals', albeit at a different level
    - However, these 'stages' provide organization towards the utilization of techniques and tools to achieve analytics goals

“The Field Guide to DATA SCIENCE”
Booz/Allen/Hamilton, 2015

Data Science:

- Describe
  - Processing
    - Filtering, Imputation, Dimensionality Reduction, Normalization/Transformation
  - Aggregation
  - Enrichment
- Discover
  - Clustering
  - Regression
  - Hypothesis Testing
- Predict
  - Regression
  - Recommendation
- Advise
  - Local reasoning
  - Optimization
  - Simulation

"The Field Guide to DATA SCIENCE" Booz/Allen/Hamilton, 2015

For each of the most indented item, data analytics techniques are provided based on specific situations… e.g. ….:

- Describe
  - Processing
    - Filtering, Imputation, Dimensionality Reduction, Normalization/Transformation e.g., Outlier Removal, Random Sampling, K-means clustering, Fast Fourier Transformation
  - Aggregation e.g., Distribution Fitting
  - Enrichment e.g., Annotation
- Discover … and so on
  - Clustering
  - Regression
  - Hypothesis Testing
- Predict
  - Regression
  - Recommendation
- Advise
  - Local reasoning
  - Optimization
  - Simulation

“The Field Guide to DATA SCIENCE”
Booz/Allen/Hamilton, 2015

Also described are the different classes of techniques:

Transforming

Learning

Predictive

"The Field Guide to DATA SCIENCE" Booz/Allen/Hamilton, 2015

These classes pretty much correspond to ESDA types:

Transforming → Data Preparation, Data Reduction
Learning → Data Analysis
Predictive → Data Analysis

What we have to do:

- Review/Understand technology descriptions
- Categorize them by ESDA types
- Determine what goals they can support

Then We Went to AGU …

Analytics Session

- "Geophysical Science Data Analytics Use Case Scenarios"
- 12 Posters
- Will be acquiring additional Use Cases, in particular form Dan Crichton and David Wanik
- Analytics methodologies highlighted include: Decision Trees, Machine earning, Data Mining, Decision Tree

At the AGU …

Visited science posters to better understand research methodologies. Aka analytics used:

- Looked for presentations that discussed the co-analysis of multiple datasets
- Looked for presentations that described methodology techniques employed

- ‘Scanned’ 100’s of posters, identifying presentations (and through discussion with authors) that provide sought after information
  - 31 Atmospheric Science research projects identified
  - 12 Hydrology Science research projects identified
  - (Don’t read into the numbers, this is just as far as II got)
- Science research methodology techniques being used …

Science research methodology techniques being used (AGU findings)

- In atmospheric Research (atmosphere – comprised of gases):
  - Correlation Analysis; Bias Correlation
  - Regression Analysis; Bivariant Regression
  - Decision Tree
  - Machine Learning
  - Data Mining
  - Data Fusion
  - Computational Tools
  - Constrained Variational Analysis
  - Model Simulations
  - Ratios
  - Time Series Analysis
  - Spectral Analysis
  - Temporal Trending; Trend Analysis
  - Spatial Interpolation
  - Revised Averaging Scheme
  - Forward Modeling; Inverse Modeling
  - Radiative Transfer Model
  - Baysian Synthesis Inversion
  - Temporal Stability
  - Gaussian Distribution
  - Exponential Differentiation

Science research methodology techniques being used (AGU findings)

- In Hydrology Research (a liquid):
  - Linear Regression
  - Monte Carlo
  - Darcy Equation
  - Poisson Regression
  - Multi-variate time series analysis
  - BUDYKO formula
  - Smoothing (Gaussian)
  - Filtering (Destriping)
  - MESH Model

Shea

Framework for Putting it All Together

ESDA Goals	Data Preparation		Data Reduction		Data Analysis	
	ESDA Requirements	ESDA Tools/ Techniques	ESDA Requirements	ESDA Tools/ Techniques	ESDA Requirements	ESDA Tools/ Techniques
1.To calibrate data						
2.To validate data (note it does not have to be via data intercomparison)						
3.To assess data quality						
4.To perform coarse data preparation (e.g., subsetting data, mining data, transforming data, recovering data)						
5.To intercompare datasets (i.e., any data intercomparison; Could be used to better define validation/quality)						
6.To tease out information from data						
7.To glean knowledge from data and information						
8.To forecast/predict/model phenomena (i.e., Special kind of conclusion)						
9.To derive conclusions (i.e., that do not easily fall into another type)						
10.To derive new analytics tools						

Framework for Putting it All Together

ESDA Goals	Data Preparation		Data Reduction		Data Analysis	
	ESDA Requirements	ESDA Tools/ Techniques	ESDA Requirements	ESDA Tools/ Techniques	ESDA Requirements	ESDA Tools/ Techniques
1.To calibrate data	**Ingest from various sources**				**High speed processing; Math functions**	
2.To validate data (note it does not have to be via data intercomparison)	**Ingest from various sources; Homogenize data**		**Sampling**		**Visualization; Gridding**	
3.To assess data quality	**Access large datasets**				**Assess erroneous data; Detect data anomalies**	
4.To perform coarse data preparation (e.g., subsetting data, mining data, transforming data, recovering data)	**Access large datasets**		**Subsetting, mining, machine learning**		**High speed processing**	
5.To intercompare datasets (i.e., any data intercomparison; Could be used to better define validation/quality)	**Homogenize data**				**Intercomparison statistics; Pattern recognition**	
6.To tease out information from data	**Seek heterogeneous data relationships; Ingest from various sources**				**Seek data relationships; Image processing**	
7.To glean knowledge from data and information	***Looking***	***for***	***Community***	***input***		
8.To forecast/predict/model phenomena (i.e., Special kind of conclusion)	**Data exploration; Near Real Time data**		**Neural networks**		**Math/Stat modeling**	
9.To derive conclusions (i.e., that do not easily fall into another type)	**Data exploration; code**		**Filter, mine, fuse, interpolate data**		**Manage custom code**	
10.To derive new analytics tools	**Access very large datasets; homogenize data**				**Visualization**	

Are We on the Right Track?

- √ Derive requirements from use cases (still needs work) *
- - Further validate requirements with (many) more additional use cases
- √ Survey/Describe existing data analytics tools/techniques *
- - Perform gap analysis between ESDA requirements and available tools *

More Use Cases

Looking for more use cases…..

Thank you

BACKUP

NIST Big Data Definitions and Taxonomies, V 0.9

National Institute of Standards and Technology (NIST) Big Data Working Group (NBD-WG)
February, 2014, http://bigdatawg.nist.gov/show_InputDoc.php, M0142

Big Data *consists of extensive datasets, primarily in the characteristics of* ***volume***, ***velocity*** *and/or* ***variety***, *that require a scalable architecture for efficient storage, manipulation, and analysis.*

Open Geospatial Consortium (OGC)
Big Data Working Group

http://external.opengeospatial.org/twiki_public/BigDataDwg/WebHome

*"**Big Data**" is an umbrella term coined by Doug McLaney and IBM several years ago to denote data posing problems, summarized as the **four Vs**:*

- ***Volume** – the sheer size of "data at rest"*
- ***Velocity** – the speed of new data arriving ("data at move")*
- ***Variety** – the manifold different*
- ***Veracity** – trustworthiness and issues of provenance*

IEEE BigData 2014

http://cci.drexel.edu/bigdata/bigdata2014/callforpaper.htm

*… in any aspect of **Big Data** with emphasis on **5Vs** (**Volume, Velocity, Variety, Value and Veracity**) relevant to variety of data (scientific and engineering, social, …) that contribute to the Big Data challenges*

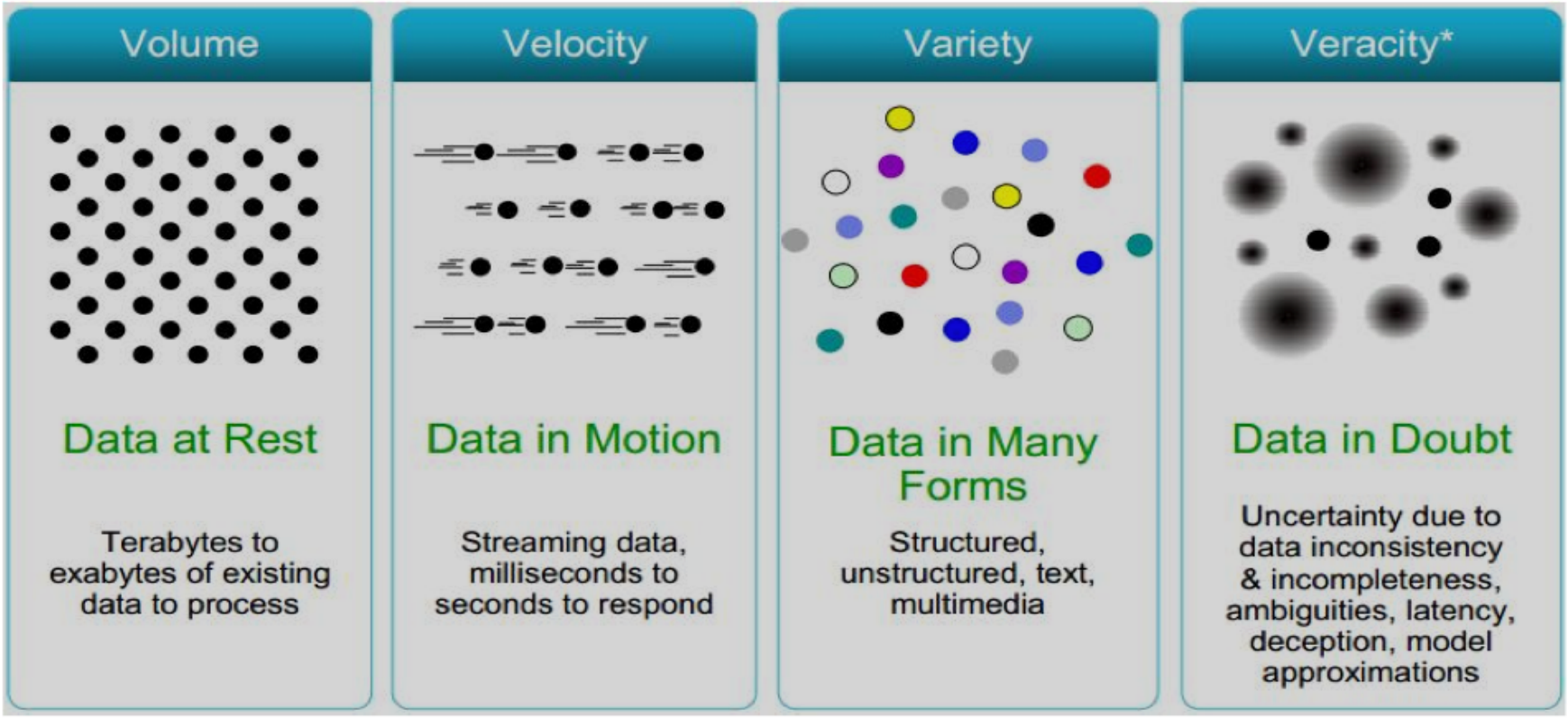
Ruth adds:

Visibility

From: Demystifying Data Science

(Natasha Balac , accessible via: http://bigdatawg.nist.gov/show_InputDoc.php, M0169)

4 V's of Big Data



IBM, 2012

So, Why does Big Data Have Everybody's Attention?

This is an encourager: (http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf)

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges.

Data Scientist in the context of analytics

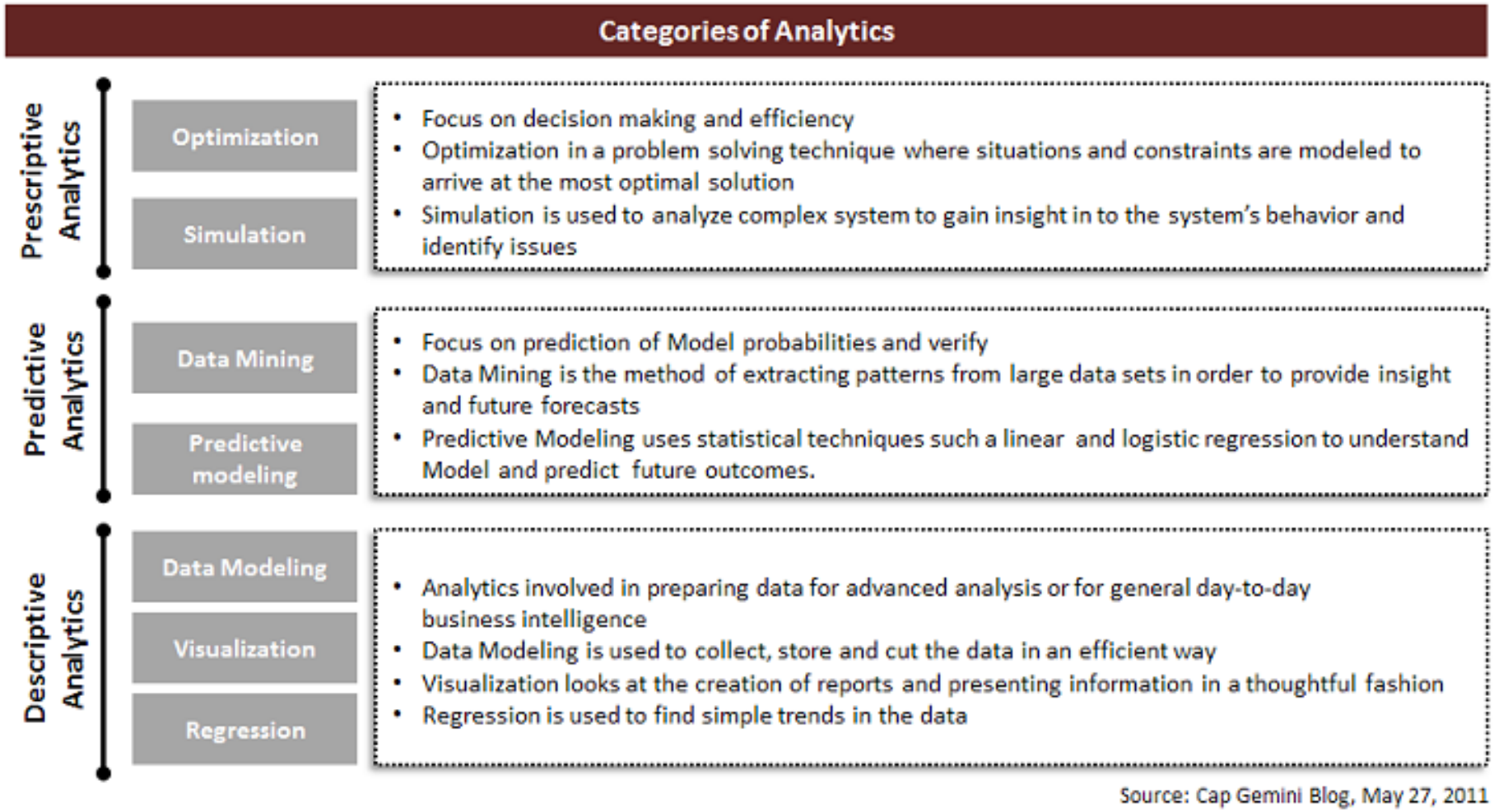
Data Scientist

A data scientist possesses a combination of analytic, machine learning, data mining and statistical skills as well as experience with algorithms and statistical skills as well as experience with algorithms and coding. Perhaps the most important skill a data scientist possesses, however, is the ability to explain the significance of data in a way that can be easily understood by others. (Source: http://searchbusinessanalytics.techtarget.com/definition/Data-scientist)

Rising alongside the relatively new technology of big data is the new job title data scientist. **While not tied exclusively to big data** projects, the data scientist role does complement them because of the increased breadth and depth of data being examined, as compared to traditional roles. (Source: http://www-01.ibm.com/software/data/infosphere/data-scientist/)

Analytics

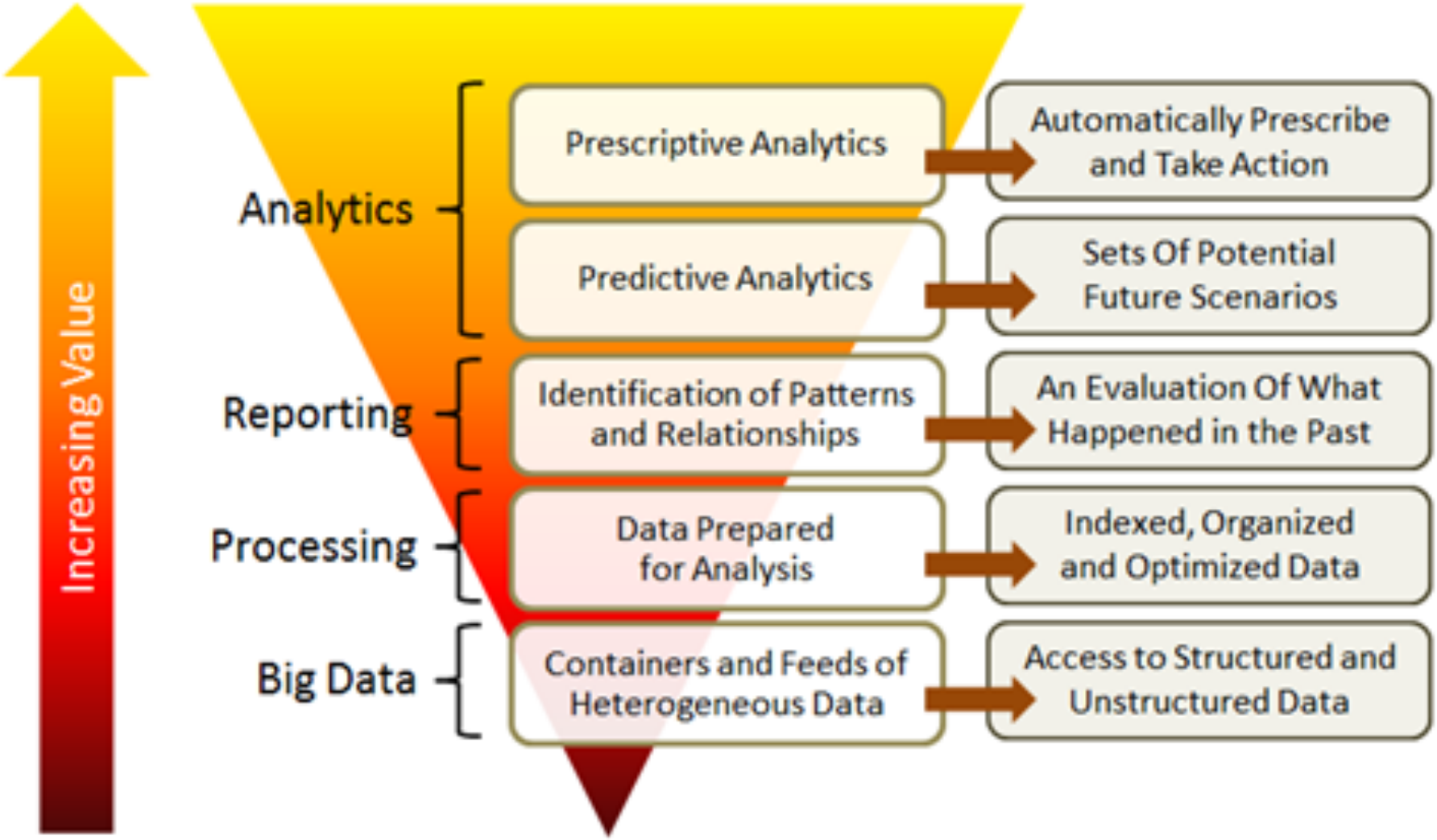
(http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/)



Source: Cap Gemini Blog, May 27, 2011

Another look at Analytics

(http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/)



2014 IEEE International Conference on Big Data (IEEE BigData 2014) Call for papers in the following (consolidated) areas:	What V's do the call for papers address:			
	Volume	**Velocity**	**Variety**	**Veracity**
1. Big Data Science and Foundations				
a. Novel Theoretical Models for Big Data	√	√	√	√
b. New Computational Models for Big Data	√	√		
c. Data and Information Quality for Big Data	√			√
d. New Data Standards			√	√
2. Big Data Infrastructure				
a. High Performance/Parallel/Cloud/Grid/Stream Computing for Big Data	√	√		
b. Autonomic Computing and Cyber-infrastructure, System Architectures, Design and Deployment	√	√		
c. Programming Models, Techniques, and Environments for Cluster, Cloud, and Grid Computing to Support Big Data	√			
d. Big Data Open Platforms	√	√		
e. New Programming Models and Software Systems for Big Data beyond Hadoop/MapReduce, STORM	√		√	

2014 IEEE International Conference on Big Data (IEEE BigData 2014) Call for papers in the following (consolidated) areas:	What V's do the call for papers address:			
	Volume	**Velocity**	**Variety**	**Veracity**
3. Big Data Management				
a. Algorithms, Architectures, and Systems for Big Data Web Search and Mining of variety of data.		√	√	√
b. Algorithms, Architectures, and Systems for Big Data Distributed Search	√	√	√	
c. Data Acquisition, Integration, Cleaning, and Best Practices			√	√
d. Visualization Analytics for Big Data			√	√
e. Computational Modeling and Data Integration	√		√	
f. Large-scale Recommendation Systems and Social Media Systems			√	√
g. Cloud/Grid/Stream (Semantic-based) Data Mining and Pre-processing- Big Velocity Data		√	√	√
h. Multimedia and Multi-structured Data- Big Variety Data			√	

A 2011 McKinsey report suggests suitable technologies include...

(http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

…A/B testing, association rule learning, classification, cluster analysis, crowdsourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, anomaly detection, predictive modelling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualisation.

Analytics Master's Degrees Programs